Analysis of Patent Databases Using VxInsight™

Kevin W. Boyack†, Brian N. Wylie, George S. Davidson, David K. Johnson

Sandia National Laboratories*
Albuquerque, NM 87185 USA
† 505-844-7556

† kboyack@sandia.gov

ABSTRACT

We present the application of a new knowledge visualization tool, VxInsight, to the mapping and analysis of patent databases. Patent data are mined and placed in a database, relationships between the patents are identified, primarily using the citation and classification structures, then the patents are clustered using a proprietary force-directed placement algorithm. Related patents cluster together to produce a 3-D landscape view of the tens of thousands of patents. The user can navigate the landscape by zooming into or out of regions of interest. Querying the underlying database places a colored marker on each patent matching the query. Automatically generated labels, showing landscape content, update continually upon zooming. Optionally, citation links between patents may be shown on the landscape. The combination of these features enables powerful analyses of patent databases.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, relevance feedback*

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *graphical user interface, interaction styles, screen design*

I.5.3 [**Pattern Recognition**]: Clustering – *similarity measures*

J.1 [**Administrative Data Processing**]: Law

General Terms

Algorithms, Management, Economics, Human Factors.

Keywords

Patent analysis, 3-D visualization, interactive navigation, data mining, clustering.

1. INTRODUCTION

Our ability to collect data vastly exceeds our ability to make sense of it. Commercial databases, the World Wide Web and data warehouses offer the promise of insight to those who can extract information from the overwhelming volume of raw data. A number of data mining tools exist to answer specific questions about data collections. However, more often than not, an analyst doesn't even know what questions to ask, and large-scale trends can be more important than narrow observations. Unfortunately, existing tools are not well suited to interact with data in an intuitive exploratory manner.

At Sandia National Laboratories, we face the same challenges in the information arena as does most of industry. In particular, we needed to answer the question, "Where should we put our next research dollar?" Our response to this question was to develop a new software tool, VxInsight™, to allow us to build maps of technology using data from the Science Citation Index [4]. Over the past few years, we have found that VxInsight has broad application to mapping and navigation of many different types of data. The application of the present paper is the mapping and analysis of patent data.

VxInsight™ is a powerful and flexible tool for exploring data collections. It works by providing access to the data in an intuitive visual format that is easy to interpret and which aids natural navigation. Millions of years of evolution have equipped us with extraordinary powers, within our visual cortex, to spot trends and patterns, to identify outliers and to detect relationships. VxInsight exploits this capability by presenting the data as a landscape, a familiar representation that we are adept at interpreting, and which allows very large data sets to be represented in a memorable way.

2. RELATED WORK

2.1 Patent analyses

The majority of patent analyses represent proprietary work, and are thus not available in the open literature. These analyses are most often done in-house (for instance, by a large company evaluating their intellectual property portfolio), or contracted out to a firm specializing in this service. Additionally, some firms provide patent analyses specific to certain industries, which are then available for purchase by any interested party.

Despite the proprietary nature of this work, some clues exist in the literature as to the nature of many patent analyses. First, there is the patent search which may be done on-line or through a variety of paid databases vendors (e.g. Lexis-Nexis). See, for example

Presented at *New Paradigms in Information Visualization and Manipulation*, a Workshop at the *9th International Conference on Information and Knowledge Management (CIKM 2000)*, November 10, 2000, McLean, VA.

* **Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.**

[13], which is a recent review of patent search strategies. More detailed analyses are generally indicator-based, i.e. they use indicators such as number of patents, number of citations, etc. segmented by assignee, classification, and year to show patent positions and trends. Such studies have been done to show global trends [1], trends in a particular discipline [6], or to inform policy [10]. Recently, some analysts have begun using co-citation analysis [11] to cluster patents. Clustering of patents using text-based measures is also available using some commercial tools (e.g. Delphion, Aurigin).

2.2 Literature maps

Mapping of patent data is very similar to the mapping of literature data, which has been done for many years. Various efforts to map the structure of science from literature have been undertaken. The majority of these studies are performed at the discipline or specialty level. Maps are typically based on similarity between journal articles using citation analysis [14], co-occurrence or co-classification using keywords, topics, or classification schemes [12, 16], or journal citation patterns [9]. Many of these studies probe the dynamic nature of science, and the implications of the changes.

Once a similarity matrix is defined, algorithms are used to cluster the data objects (e.g., articles or patents). Common clustering methods for producing maps include multidimensional scaling, hierarchical clustering, k-means algorithms, and self-organizing maps. The standard mapping output for the literature studies referenced above is a circle plot where each cluster is represented by an appropriately sized circle. Links between circles provide relationship information. Traditionally, map outputs have been paper-based and only resolve structure at a few discrete levels. However, in recent years, several systems have been reported that use a computer display and allow some navigation of the map space.

2.3 Visualization tools

SENTINEL [5] is a Harris Corporation package that combines a retrieval engine using n-grams and context vectors for effective query with a visualization system called VisualEyes™. The visualization tool allows the user to interact with document clusters in a three-dimensional space. Chen [3] uses a VRML 2.0 viewer in conjunction with Generalized Similarity Analysis to display papers (as spheres) and the Pathfinder linkage network connecting them which has been calculated from a co-citation analysis.

Self-organizing maps have been used in many venues, including the organization of documents [8]. These maps are used to position documents, and then display them in a two-dimensional contour-map-like display in which color represents density. Peak labels can be generated automatically, and some limited navigational and retrieval capabilities are provided.

Two packages that are similar to VxInsight are SCI-Map developed by ISI [15], and the SPIRE suite of tools, which originated at Pacific Northwest National Laboratory [7, 17]. SCI-Map uses a hierarchically nested set of maps to display the document space at varying levels of detail. This nesting of maps allows drilling-down to subsequent levels. Each map is similar to the traditional circle plot, where the size of the circle can indicate the density of documents contained in the circle, or some other measure of importance.

Like VxInsight, SPIRE maps objects to a two-dimensional plane so that related objects are near each other, and provides tools to interact with the data. SPIRE has two visualization approaches. In the Galaxies view, documents are displayed as a scatter plot. This interface allows drilling down to smaller sections of the scatter plot, and provides some summarization tools. In the Themescape view, a high-level terrain display, similar to that in VxInsight, is used. Themescape visualizes specific themes as mountains and valleys, where the height of a mountain represents the strength of the theme in the document set. However, unlike VxInsight, this view is static, and allows no zooming into or out of the terrain to get more detailed information. Thus, SPIRE does not provide the continuous, multi-resolution viewing that is essential for revealing the inherent structure of an information space at multiple levels of detail.

3. SOFTWARE DESCRIPTION

VxInsight displays a set of abstract objects (e.g. scientific documents, patents, financial transactions) using geometric proximity to convey similarity. The more similar two objects are, the closer together they will be placed on the landscape. The analyst has the ability to use standard similarity functions or to define a custom function. Using these similarity values, each object is assigned a location on a 2D plane. For example, with patents, citation analysis can be used as a measure of similarity. The more citations two patents share, the higher their similarity, and the closer they will be on the landscape.

The landscape is displayed 'on the fly' with the height of each mountain being proportional to the number of objects beneath it. Labels for peaks are also generated on the fly, revealing the content of the objects that comprise the mountain. The tool supports multi-resolution zooming into the landscape to explore interesting regions in greater detail, which reveals structure on multiple scales. Following each mouse click, the landscape is recalculated, to give a new, higher resolution view of the desired terrain. Temporal data can be viewed using a time slider to reveal growth and reduction in areas of interest, new emerging areas, and bridged regions that have merged together.

Data access and retrieval is achieved via an ODBC connection to the user's database. VxInsight uses the ODBC connection in conjunction with the Structured Query Language (SQL) to provide the user with an intuitive and powerful interface. Clicking on objects invokes specific detailed information (such as patent title), interrogating the database lights up the landscape showing query matches. Competitive analysis tasks can be accomplished by combining multiple queries for simultaneous comparison. A video is available online (http://www.sandia.gov/VxInsight) demonstrating a number of these features.

4. APPLICATION TO PATENT DATA

An obvious application of VxInsight is in analysis of patent data to build a detailed map of a specific intellectual property category. This may occur as part of a roadmapping event [2], or as a management or legal exercise. Questions that one may ask include:

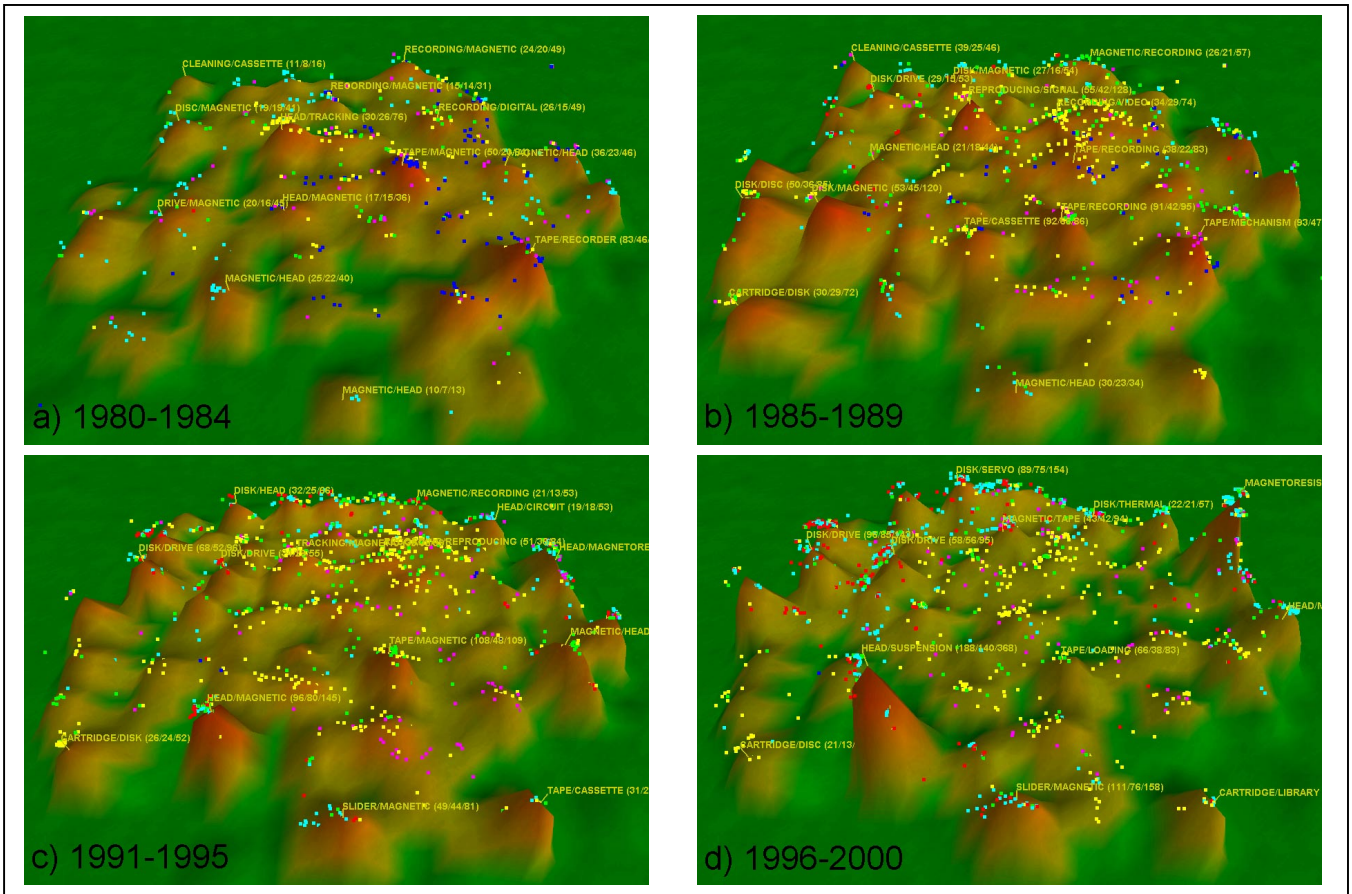

Figure 1. Landscapes of patent class 360 for four different five-year time periods.
(Company color legend – blue: Olympus, yellow: Sony, green: Hitachi, magenta: Philips, cyan: IBM, red: Seagate)

- What are the logical groupings of patents?
- What are the potential overlaps between groupings?
- What are the trends?
- What other patents are related to mine?
- Are there holes in my portfolio that need to be filled?
- Has someone else already done what I want to do?
- Can the technology I need be licensed from another company?
- Are there other companies that might be interested in my technology?

Information that will aid in answering these types of questions can be gleaned from maps of intellectual property based on patent data.

Preparation of patent data for analysis with VxInsight requires several general steps:

1. procurement of the data by query or otherwise,
2. calculation of a similarity measure based on citations or common content,
3. ordination or clustering.

Clustering may be done using any algorithm available to the user. However, we use a force-directed placement algorithm [4] that resides in the VxInsight software environment to create patent maps.

In this study we demonstrate two different subsets of the US patent database – one based on a particular technology, denoted by classification, and one showing all the patents issued during a one month time window. Each patent set was generated by query of the US Patent bibliographic file (front page) information that we have placed in a MySQL database. Patent bibliographic file data are available from the US Patent and Trademark Office.

4.1 Patents in Class 360

Query of the US Patent database for all patents whose primary classification is class 360 (Dynamic Magnetic Information Storage or Retrieval) returned 15,782 patents over the time period from 1976 to September 2000. These patents are segmented into over 200 different subclasses. 55,553 citations link all but 780 of these patents. This does not include any citations to or from patents outside the 360 class.

A similarity measure was calculated using the direct and co-citation link types of Small [14]. Direct citations were given a weighting five times that of each co-citation link. Ordination was done using the VxOrd force-directed placement algorithm within VxInsight.

Maps showing the patent class 360 landscape for four different time periods are shown in Figure 1. Class 360 is dominated by

disk drive and tape (audio and video) drive technologies. Tape-related technologies occur in the Figure 1 landscapes in the middle section outward to the edge of the landscape at the 4:00 position. Disk-related technologies are found in most of the rest of the landscape, particularly to the left and at the edges.

The progression in disk and tape drive technologies at the macro level are shown by the sequence of maps in Figure 1. Tape-related technologies dominate in the early 1980's (the peaks at the center and 4:00 position in Figure 1a are larger than the other peaks), but are surpassed by disk-related technologies with the progression of time. One of the major players in tape technologies in the early 1980's, Olympus Optical, disappears from the scene entirely in later years. Sony and Hitachi produce patents in the same tape-related areas (much of it for video) throughout the 1980's and 1990's. However, Philips, a primary competitor to Sony, has most of its patents in different parts of the landscape than Sony.

The dominance of disk-related technologies becomes very evident by the late 1990's with magnetoresistive heads (2:00 position in Figure 1d) and disk heads with suspension mechanisms (7:00 position in Figure 1d) leading the way. As disk-related patents become more prevalent in the 1990's, the relative patenting of firms such as IBM and Seagate is easily seen.

Changes in focus of single peaks can be noted by watching the labels change with time. For example, the magnetic head peak at the 6:00 position shows an evolution to head slider technology in the 1990's.

Similar analyses can be done at the micro level. For instance, a query for all patents assigned to the Eastman Kodak Corporation returns 225 patents and shows their position on the landscape. These patents show up in many areas, but there are several concentrated clusters of patent activity. One of these clusters is shown in Figure 2, where the Kodak patents are colored and have been labeled (outside of VxInsight) with letters in the order of their issue dates. A listing of these patents, their issue dates, titles, and the number of times each patent has been cited, is given in Table 1.

One of the relative strengths of the VxInsight process for clustering, displaying, and navigating patent sets is shown by comparing the data in Figures 2 and 3. Figure 3 contains a link diagram of the Kodak patents from Table 1. Figures 2 and 3 both clearly show that there are no citation links between the first five patents in the group. Table 1 shows that these are all highly cited (minimum of 14 citations) patents, and thus are important.

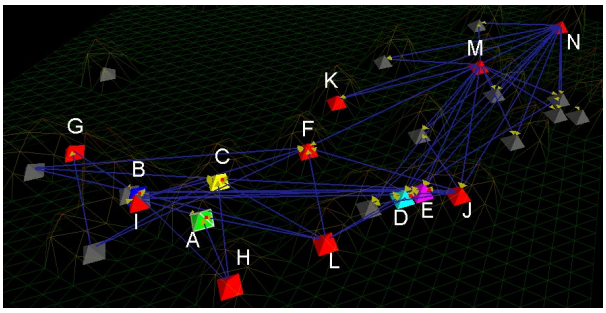

Figure 2. Eastman Kodak patents (colored pyramids) from the patent class 360 map related to magnetic recording of information on film.

Table 1. Listing of Eastman Kodak patents shown in Figure 2.

	Patent No.	Issue Date	N	Title
A	4,933,780	6-12-90	26	Camera apparatus for magnetically recording on film
B	5,034,836	7-23-91	23	Magnetic head suspension apparatus for use with a photographic film
C	5,041,933	8-20-91	14	Magnetic head suspension apparatus for use with a photographic film
D	5,274,522	12-28-93	14	Magnetic head-to-media backer device
E	5,285,324	2-8-94	15	Magnetic head-to-recording medium support apparatus
F	5,285,325	2-8-94	12	Web guiding device for use in a magnetic reading and/or recording apparatus
G	5,400,200	3-21-95	2	Magnetic head suspension apparatus
H	5,535,062	7-9-96	4	Magnetic reading and/or recording apparatus for reading and/or recording information on a magnetic information track on a photosensitive medium
I	5,563,751	10-8-96	0	Longitudinal bending interface for thick film magnetic recording
J	5,576,916	11-19-96	3	Magnetic head-to-media backer device
K	5,598,310	1-28-97	3	Magnetic head-to-media backer assembly
L	5,721,652	2-24-98	0	Roll stabilized, nesting vee, magnetic head assembly for magnetics-on-film
M	5,764,456	6-9-98	0	Apparatus for backing a magnetic medium in contact with a magnetic read/write head
N	5.923,507	7-13-99	0	Magnetic head-to-medium backer device

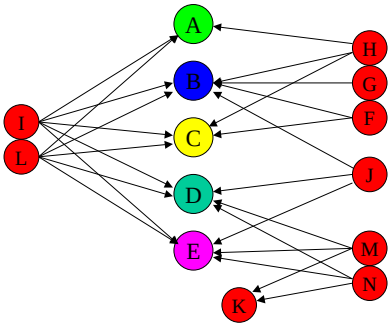

Figure 3. Link diagram of the Kodak patents shown in Figure 2 and Table 1.

If one were performing a traditional link analysis (showing direct references and/or citations to a single patent) based on any one of the first five patents in this group, the other four important patents would not be found. Thus, any analysis would be biased by lack of knowledge of how these patents fit in the technology. *By contrast, VxInsight places these patents close to each other*

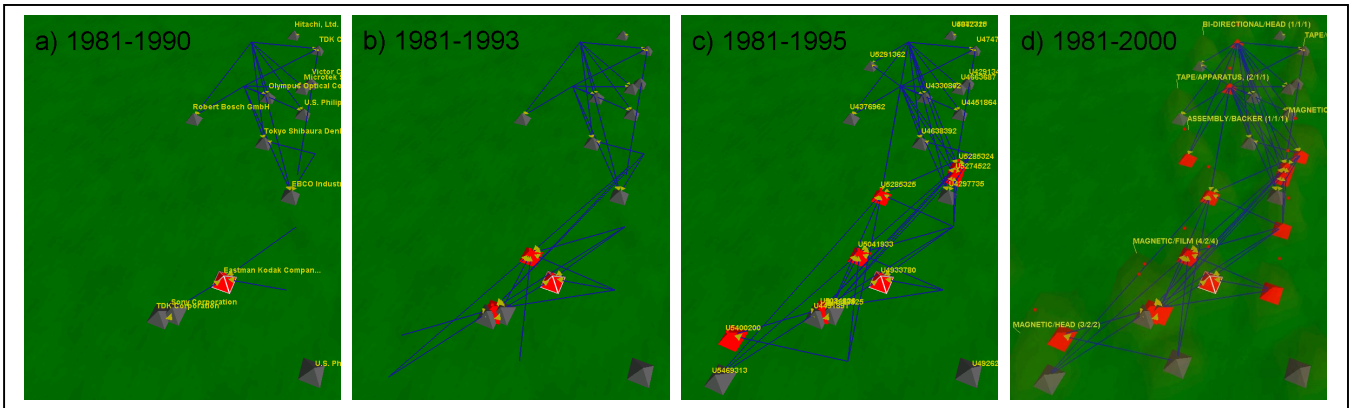

Figure 4. Evolution of the group of patents regarding magnetic recording of information on film from the patent class 360 map.

because of the other patents that commonly cite them. For example, patents A-C are all cited by patents H, I, and L, and thus are placed close to each other on the map. Likewise, patents D and E are both cited by patents I, J, L, M, and N, and so appear next to each other in the map.

VxInsight also allows the analyst to study the evolution of the technology over time. Figure 4 shows the same group of patents at four different time steps, where the growth of the field is plain to see. VxInsight provides context to groups of patents that allows accurate and powerful analyses to be performed.

4.2 Patents issued in January 2000

The US Patent Office issued a total of 10,805 patents during the month of January 2000. These patents were spread over many of the hundreds of patent classes. Given that all of these patents issued during the same month, there are no citations between patents in this set. Thus, the current classification (primary and additional classifications) for each patent was used as the basis for similarity between patents. Patents in the same class and subclass were given a similarity weighting five times that of patents in the same class, but different subclasses. Patents with two classifications in common were weighted doubly, and so forth.

One might assume that when clustering patents using similarities derived from the classification system, the patents would simply divide by primary class. However, this is not the case. (One exception will be shown later.) Since patents are often classified into two or more classes, the similarity includes information about the relationship between classes and subclasses. The resulting clustering using the VxInsight ordination algorithm shows a much richer segmentation than a simple division by class.

In the January 2000 patent map shown in Figure 5, computer and software related patents populate the top of the map, patents in chemistry, biology, and medicine occur at the right and lower right, and mechanical patents in a very broad sense of mechanical fill the middle of the map.

A query for all design patents (magenta dots in Figure 5) indicates that nearly all of the design patents lie in well-defined clusters at the left and bottom edges of the landscape. This may seem curious at first. However, a review of the classification of the design patents reveals that very few design patents have additional classifications outside their specific design class. Thus, the similarity measure contains very little information linking the design classifications to the non-design classifications, and the design clusters move to the edges in the ordination step.

There are some exceptions to this observation. For example, the peak labeled INK/PRINTER near the right side of the landscape in Figure 5 contains some design patents. A closer view of the PRINTER cluster shows that it contains 18 design patents (class D18) and 87 non-design patents (of which 60 are in class 347). One would think that the 18 patents from class D18 would have formed their own cluster and moved to the edge of the map. However, two of the D18 patents are also classified in class 347. This very small linkage – (2 of 18) linked to (2 of 60) – pulls the two groups together into one cluster. Thus, printer technology and printer design (i.e., look rather than function) lie together in the VxInsight map.

Many other analyses, both at the macro and micro levels, can be illustrated from this set of patents. However, we will make only one more observation, based on Figure 5. Patents granted to universities are shown as green dots. They occur primarily in the areas of chemical compositions and genomics research (the two large peaks at the right without labels). Obviously, a great deal of fundamental research is done at universities throughout the US in all technical fields. It is interesting, therefore, that the majority of patents granted to universities recently are in two well-defined areas, and suggests that the intellectual property departments of universities may see a strong economic future in genomics and chemical compounds.

5. CONCLUSIONS AND FUTURE WORK

The example analyses above show that VxInsight is a powerful tool to aid in analysis of patent data. VxInsight provides context that is difficult to find using other analysis tools and techniques, and which can enable more accurate analyses. The following specific conclusions about the use of VxInsight with patent data can be supported as well:

- Use of a similarity measure based on citation links works well for a patent set centered within a particular patent class.
- Use of a similarity measure based on current classification works well for a patent set with no citation structure, provided the set encompasses many different patent classes.

In the future we plan enhancements to VxInsight to allow more robust analyses. These enhancements include turning on citation links selectively rather than all at once, calculation of indicators

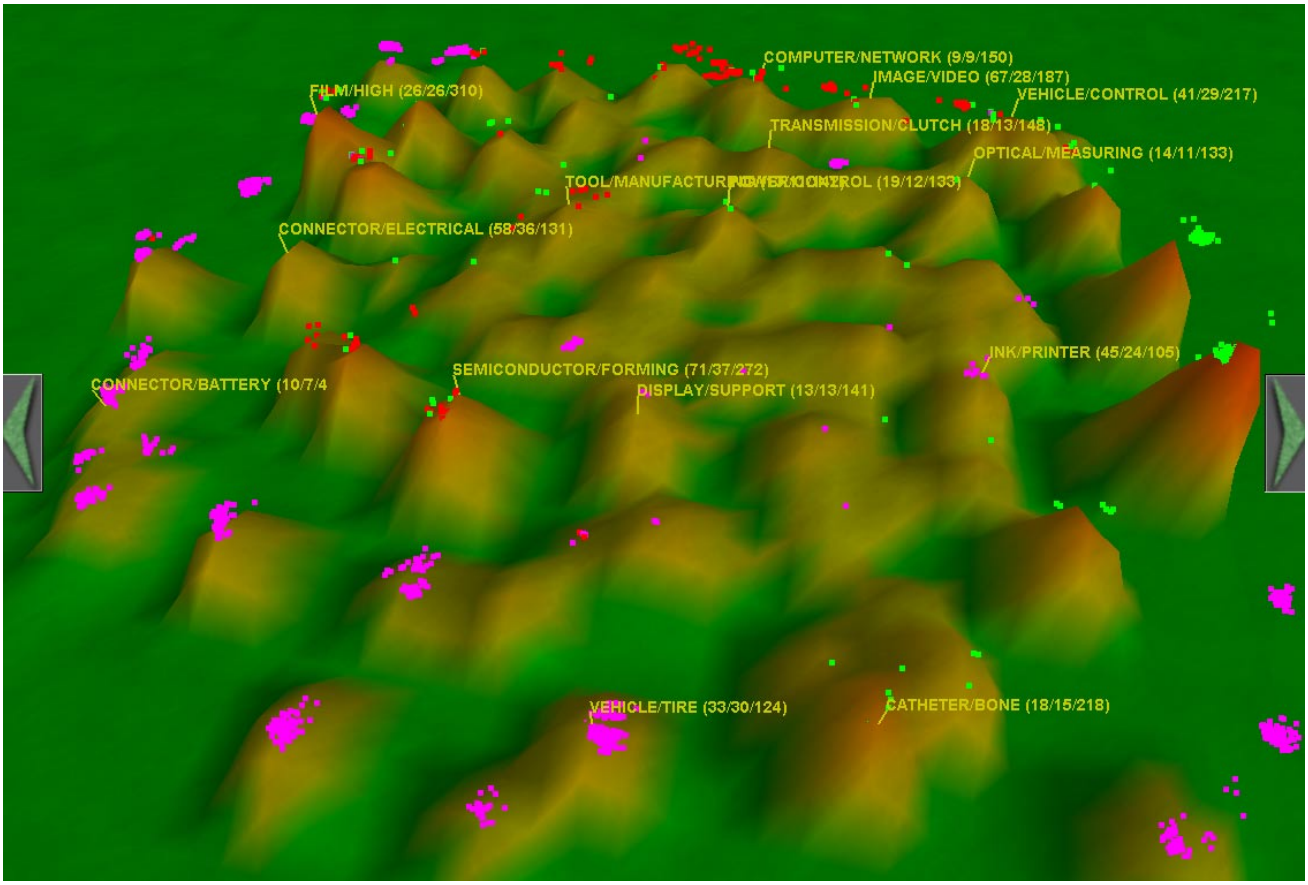

Figure 5. Map of all patents issued by the US Patent Office in January 2000.
(Color legend - magenta: design patents, green: patents granted to universities, red: IBM)

so that the indicators can be more tightly coupled with context, and enabling a web link to any specific patent at the US Patent Office web site from VxInsight. We also plan to investigate the effect of coupling citation and classification-based similarities for patent data.

6. REFERENCES

[1] Albert, M.B., Yoshida, P.G., and van Opstal, D. Global patenting trends. CHEMTECH 29(2), 47-58 (1999).

[2] Beck, D. F., Boyack, K. W., Bray, O. H. & Siemens, W. D. Landscapes, games, and maps for technology planning. CHEMTECH 29(6), 8-16 (1999).

[3] Chen, C. Visualising semantic spaces and author co-citation networks in digital libraries. Information Processing and Management 35, 401-420 (1999).

[4] Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E. & Wylie, B. N. Knowledge mining with VxInsight: discovery through interaction. Journal of Intelligent Information Systems 11, 259-285 (1998).

[5] Fox, K. L., Frieder, O., Knepper, M. M. & Snowberg, E. J. SENTINEL: A multiple engine information retrieval and visualization system. Journal of the American Society for Information Science 50(7), 616-625 (1999).

[6] Gupta, V.K. & Pangannaya, N.B. Carbon nanotubes: bibliometric analysis of patents. World Patent Information 22, 185-189 (2000).

[7] Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. Multi-faceted insight through interoperable visual information analysis paradigms. Proceedings of IEEE Information Visualization '98, 137-144 (1998).

[8] Honkela, T., Kaski, S., Kohonen, T. & Lagus, K. Self-organizing maps of very large document collections: Justification for the WEBSOM method. In I. Balderjahn, R. Mathar & M. Schader (Eds.) *Classification, Data Analysis, and Data Highways.* Berlin: Springer (1998).

[9] Leydesdorff, L. The generation of aggregated journal-journal citation maps on the basis of the CD-ROM version of the Science Citation Index. Scientometrics 31, 59-84 (1994).

[10] Margolis R.M. & Kammen, D.M. Evidence of under-investment in energy R&D in the United States and the impact of Federal policy. Energy Policy 27, 575-584 (1999).

[11] Mogee, M.E. Patent analysis methods in support of licensing. Presented at the Technology Transfer Society Annual Conference (July 22, 1997, Denver, CO). http://www.mogee.com/reports/methods.html.

[12] Noyons, E. C. M. & Van Raan, A. F. J. Advanced mapping of science and technology. Scientometrics 41, 61-67 (1998).

[13] Schwander, P. An evaluation of patent searching resources: comparing the professional and free on-line databases. World Patent Information 22, 147-165 (2000).

[14] Small, H. Update on science mapping: creating large document spaces. Scientometrics 38, 275-293 (1997).

[15] Small, H. Visualizing science by citation mapping. Journal of the American Society for Information Science 50(9), 799-813 (1999).

[16] Spasser, M. A. Mapping the terrain of pharmacy: co-classification analysis of the International Pharmaceutical Abstracts database. Scientometrics 39, 77-97 (1997).

[17] Wise, J. A. The ecological approach to text visualization. Journal of the American Society for Information Science 50(13), 1224-1233 (1999).